Safe Functional Inference for Uncharacterized Viral Proteins

Yaniv Loewenstein[1], Michal Linial[2*]
[1]School of Computer Sciences and Engineering, [2]The Sudarsky Center for Computational Biology, Institute of Life Sciences, The Hebrew University of Jerusalem, 91904, Israel
*To whom correspondence should be addressed: michall@cc.huji.ac.il

1. INTRODUCTION

The explosive growth in the number of sequenced genomes has created a flood of protein sequences with unknown structure and function. A routine protocol for functional inference on an input query sequence is based on a database search for homologues. Searching a query against a non-redundant database using BLAST (or more advanced methods, e.g. PSI-BLAST) suffers from several drawbacks: (i) a local alignment often dominates the results; (ii) the reported statistical score (i.e. E-value) is often misleading; (iii) incorrect annotations may be falsely propagated.

Several systematic methods are commonly used to assign sequences with functions on a genomic scale. In Pfam (1) and resources alike, statistical profiles (HMMs) are built from semi-manual multiple alignments of seed homologous sequences. The profiles are then used to scan genomic sequences for additional family members. The drawbacks of this scheme are: (i) only families with a predetermined seed are considered; (ii) the query must have a detectable sequence similarity to seed sequences; (iii) attention to internal relationships among the family members or the relations to other families is lacking; (iv) family membership is often set by pre-determined thresholds.

An alternative to profile or model based methods for functional inference relies on a hierarchical clustering of the protein space, as implemented in the ProtoNet approach (2). The fundamental principle is the creation of a tree that captures evolutionary relatedness among protein families. The tree construction is fully automatic, and is based only on reported BLAST similarities among clustered sequences. The tree provides protein groupings in continuous evolutionary granularities, from closely related to distant superfamilies. Clusters in the ProtoNet tree show high correspondence with homologous sequence (i.e. Pfam and InterPro), functional (i.e. E.C. classification) and structural (i.e., SCOP) families (3). A new clustering scheme (4) has provided an extensive update to the ProtoNet process, which is now based on direct clustering of all detectable sequence similarities.

Herein, we use the ProtoNet resource to develop a methodology for a consistent and safe functional inference for remote families. We illustrate the success of our approach towards clusters of poorly characterized viral proteins. Viral sequences are characterized by a rapid evolutionary rate which drives viral families to be even more remote (sequence-similarity-wise). Thus, functional inference for viral families is apparently an unsolved task. Despite this inherent difficulty, the new ProtoNet tree scaffold reliably captures weak evolutionary connections for viral families, which were previously overlooked. We take advantage of this, and propose new functional assignments for viral protein families.

2. METHODS

The ProtoNet Resource: We constructed a hierarchical tree (ProtoNet5.1) for 1.8 million non-redundant (UniRef90, maximum 90% sequence identity) proteins, representing 2.5 million UniProtKB proteins. The tree construction is based on a novel algorithm (4) for clustering huge sequence data – here 1.5 billion non-trivial BLAST similarities - with a theoretical exactness guarantee. 61% of the tree sequences (UniRef90) are assigned to at least one family by Pfam (here, 8,168 families). The average size of Pfam families on the non-redundant tree sequences is 178±567. 6,882 families have at least 10 members.

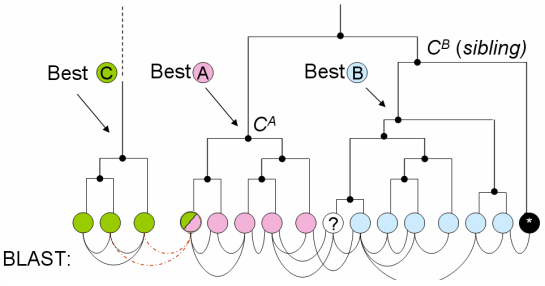

Fig. 1. *Superfamily tree search illustration*. Pink and blue represent proteins in homologous families *A* and *B*, while green and black denote other families *C* and *D*. Reported BLAST similarities are depicted by curved edges (bottom). *A* and *C* coincide on a multi-domain protein (pink and green protein) which may induce false-transitivity – falsely clustering *A* with non-homologous *C* due to local BLAST similarities of multi-domain protein (red edges). Correct merging of *A* and *B* is aided by an unassigned protein (white).

Connecting Functional Clusters: A systematic approach was developed to suggest undetected relations between homologous protein families. Briefly, we calibrate the tree for the granularity of each inspected family, and then test for other families in the same putative superfamily that is suggested by the tree. Only families which are captured well by the clustering are considered

Best Cluster & Good Siblings: The Jaccard coefficient is used as a correspondence score (*J*) of a cluster to each external assignment grouping – here, Pfam families. The "best cluster" (2) for each keyword A, is defined to be the cluster with the highest correspondence score - *J(A)*. For each keyword *A*,

where $J(A) \geq J^{A}_{cutoff}$, we have inspected the tree-sibling, i.e., the nearest cluster with whom it was merged. In those cases where the sibling cluster C^B corresponds well ($J \geq J^{B}_{cutoff}$) with another protein family keyword *B*, keywords *A* and *B* are hypothesized to be evolutionary related. Our cutoffs ensure that family relatedness is supported by relatively complete and uncontaminated families. Since our method is unsupervised, unannotated sequences guide the clustering as well, and often contribute additional support to family relatedness. Error prone inference due to multiple domain proteins is diminished, by eliminating *A* and *B* pairs with seldom coincidence on the same protein.

3. RESULTS & DISCUSSION

***Pfam Tree Correspondence and AB-pairs*:** Pfam families are very well captured by the ProtoNet tree. For 8,095 (out of 8,158) non-trivial families (≥2 non-redundant members) we have achieved on average *J*=0.893±0.175, specificity= 0.963± 0.092 and sensitivity=0.918± 0.157. Single domain, or fixed domain architectures, comprise the majority of the data, and are captured better by the tree, compared to a handful of domain families that appear in promiscuous domain architectures.

From this set, our method predicts 710 links between unique *AB*-pairs of putative homologous families (*AB*=*BA*). For this subset of Pfam families, we achieve *J*=0.93±0.09 (specificity=0.98, sensitivity=0.95) for the best clusters (*A*) and 0.88±0.12 (0.93, 0.94) for the siblings (*B*). For the reported 710 *AB*-pairs there was 14% BLAST linkage on average (E=100). For a reference, within the Pfam families the average linking is 64% at this permissive threshold.

A	B	PDB	DUF	Linkage	Number	Profile	Manual	Viral	Correct
PF04541	PF05900								True
PF04903	PF07140								True
PF05307	PF05946								P
PF05780	PF02723								True
PF06358	PF06716								True
PF05733	PF06606								True
PF06193	PF06909								True
PF06285	PF07190								True
PF06147	PF06914								True
PF03158	PF01671								P

Table 1. *Top 10 A-B clustered pairs ranked by average correspondence score.* Each category is marked by 3 levels of grey from light to dark, indicating low to high for %-linkage (<5%, 5-15%, >15%), Number of proteins in the parent cluster (<30, 30-100, >100). Similarly, color codes indicate appearance of a category in none (white), one (grey) or both paired clusters-families (dark grey) for Virus, PDB structure, DUF, manual and profile. Profile column is based on HHalign. Assign – True, when supported by several independent evidence, P - possible homology.

Characterization of a High Quality Connection – Viral Proteins: A Pfam clan unifies several Pfam families according to, e.g. external literature support, profile comparison etc. We note that within the 710 AB pairs, predictions that could be automatically validated by a clan definition, showed high accuracy with respect to the hard task of clan collection. We assigned 104 correct clan predictions vs. only 12 wrong predictions, when *A* and *B* do not coincide. Of our predictions, 366 instances had no clan assignment and provide completely new putative clan definitions. Inspection of the features that best separate correct from wrong hypotheses indicated that the average correspondence score (of *A* and *B*) should be ≥ 0.95 (we recorded only 1 false out of 48 instances, where both A and *B* had clan assignments). This subset still includes 254 of the 710 predictions, including 82 pairs containing DUF (Pfam 'domains of unknown function'). We have manually inspected 40 of the top ranked predictions (having no clan, no A-B intersection and average correspondence score ≥0.95). Viral proteins account for more than half of the top-ranked *A*-*B* pair predictions (Table 1).

As seen in Table 1, the top predictions are enriched in viral families and relatively small clusters. Many of the putative A-B pairs of viral families lead to interesting evolutionary suggestions on virus-host co-evolution as in the case of the mammalian and viral families of interferon γ–receptors (PF04903 and PF07140). The vaccinia virus interferon γ–receptor is secreted from infected cells during infection. The viral protein efficiently inhibits interferon activity leading to increased infectivity. The common evolutionary source of the two families suggests that the viral family has originated from the host proteins. Interesting biology is revealed by analyzing overlooked instances of viral proteins and DUFs. Our findings propose an automatic methodology to cluster viral proteins, and then to pinpoint hidden evolutionary connections which tie viral proteins with other protein families as well.

4. REFERENCES

1. Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L. and Bateman, A. 2008 The Pfam protein families database, *Nucleic Acids Res, 36, D281-288.*

2. Kaplan, N., Friedlich, M., Fromer, M. and Linial, M. 2004. A functional hierarchical organization of the protein sequence space, *BMC Bioinformatics, 5, 196.*

3. Kifer, I., Sasson, O. and Linial, M. 2005. Predicting fold novelty based on ProtoNet hierarchical classification, *Bioinformatics, 21, 1020-1027.*

4. Loewenstein, Y., Portugaly, E., Fromer, M. and Linial, M. 2008. Efficient algorithms for exact hierarchical clustering of huge datasets: Tackling the entire protein space. *Bioinformatics (in press).*